Maximum likelihood estimation based on Newton–Raphson iteration for the bivariate random effects model in test accuracy meta-analysis

Willis, Brian; Baragilly, Mohammed; Coomar, Dyuti

DOI:
10.1177/0962280219853602

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280219853602
journals.sagepub.com/home/smm
SAGE

Maximum likelihood estimation based on Newton–Raphson iteration for the bivariate random effects model in test accuracy meta-analysis

Brian H Willis,[1] Mohammed Baragilly[1,2] and Dyuti Coomar[1]

Abstract

A bivariate generalised linear mixed model is often used for meta-analysis of test accuracy studies. The model is complex and requires five parameters to be estimated. As there is no closed form for the likelihood function for the model, maximum likelihood estimates for the parameters have to be obtained numerically. Although generic functions have emerged which may estimate the parameters in these models, they remain opaque to many. From first principles we demonstrate how the maximum likelihood estimates for the parameters may be obtained using two methods based on Newton–Raphson iteration. The first uses the profile likelihood and the second uses the Observed Fisher Information. As convergence may depend on the proximity of the initial estimates to the global maximum, each algorithm includes a method for obtaining robust initial estimates. A simulation study was used to evaluate the algorithms and compare their performance with the generic generalised linear mixed model function *glmer* from the *lme4* package in R before applying them to two meta-analyses from the literature. In general, the two algorithms had higher convergence rates and coverage probabilities than *glmer*. Based on its performance characteristics the method of profiling is recommended for fitting the bivariate generalised linear mixed model for meta-analysis.

Keywords

Bivariate model, diagnostic accuracy, maximum likelihood estimation, meta-analysis, random effects

1 Introduction

Meta-analysis may be used to aggregate data from multiple primary studies to produce summary estimates. The most common type of model used in meta-analysis involves aggregating data where a single outcome measure is used to summarise the effect measure. Such univariate modelling approaches have yielded notable successes for meta-analysis where the results have helped inform medical decisions on treatments of life threatening diseases.[1,2]

In the case of meta-analysis of test accuracy studies, the picture is complicated by there being, in general, two outcomes of interest that are correlated. The modelling approach taken in this instance is to assume the study-level parameters for the outcomes follow a bivariate normal distribution.[3,4] Although, after a suitable transformation, we may assume the observed data within studies to be normally distributed,[3] this is an approximation and they are more accurately modelled by assuming binomial distributions.[4] Thus, to aggregate the data from test accuracy studies, a bivariate generalised linear mixed model is used. Note it is more commonly labelled a bivariate random effects model (BRM)[3] and this will be the term which will be adopted here when referring to the model.

As with many complex models of this nature, there is no closed form to the likelihood function for the model, so it is not possible to express the maximum likelihood estimates (MLEs) for the parameters analytically and numerical solutions are required. Although some packages are capable of providing maximum likelihood estimates for the parameters in the BRM, they tend to be generic packages in which the algorithms are not

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[2]Department of Applied Statistics, Helwan University, Cairo, Egypt

Corresponding author:
Brian Harvey Willis, Institute of Applied Health Research, University of Birmingam, Edgbaston, Birmingham B15 2TT, UK.
Email: b.h.willis@bham.ac.uk

readily accessible and are not necessarily optimised for this model. For example, the *glmer* function from the *lme4* package in R[5] and *NLMIXED* in SAS[6] are used to fit a range of generalised linear and non-linear models and are not specifically written for estimating the parameters in the BRM. Thus, an algorithm which is expressly written and optimised to fit the BRM has the potential for better performance characteristics than that of a generic function. It also needs to be transparent in order to facilitate understanding and reproducibility.

Here we develop two different optimisation approaches based on Newton–Raphson methods,[7] specifically to derive the maximum likelihood estimates for the parameters in the BRM. To demonstrate how this may be done from first principles, the theory and steps behind the optimisation are described explicitly, and the R code is provided in the online Appendix. We conduct a simulation study to evaluate the two algorithms and compare their performances with that of a generic function from a standard package, namely, the *glmer* function in the *lme4* package in R.[5] We then apply the algorithms to two case examples.

The paper is organised as follows. In section 2, we describe the theory in detail that underpins the bivariate random effects model used in test accuracy meta-analyses. In section 3, the optimisation methods in generic packages that may be used to fit the BRM are described. In section 4, the theory behind deriving maximum likelihood estimates in the BRM is explained in detail. In sections 5 and 6, the method of Profiling[8] and the Observed Fisher Information using robust initial parameter values (OFIRIV) are developed for the BRM. In section 7, these methods are compared using a simulation study and applying them to two case examples from the literature. Finally, in section 8, we end with the discussion.

2 Statistical methodology

A test's performance is traditionally summarised in terms of its sensitivity (the proportion of patients with disease who test positive) and specificity (the proportion of patients without disease who test negative). The two are also correlated being affected by the position of the threshold for a positive test result: as the threshold increases, the sensitivity decreases and the specificity increases. This effect is summarised by a receiver operating characteristic (ROC) curve which plots the different sensitivity–specificity pairs for each test threshold.[9]

An early attempt to incorporate such an effect in meta-analysis was made by Moses and colleagues,[10] who produced a Summary ROC (SROC) curve using simple linear regression. The model does capture variation between studies due to a changing threshold but other sources of variation are largely ignored. For the purpose of translation into practice, a summary point is usually more desirable but a valid point estimate is not readily provided by this model.

Attempts to overcome these limitations[3,4] have led to the proposing of hierarchical models.[3,4,11] Van Houwelingen[12] applied a bivariate random effects model to meta-analysis which was later taken up by Reitsma,[3] who applied it to test accuracy meta-analyses. This model allows a summary point for the sensitivity and specificity in ROC space to be estimated. An alternative approach as proposed by Rutter and Gatsonis[9] leads to a Hierarchical Summary Receiver Operating Characteristic (HSROC) curve, although a summary point may be derived from this model. Here we will focus on the bivariate random effects model for test accuracy studies. The model is a mixed model and assumes a bivariate normal distribution of the form

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}\right) \tag{1}$$

where α_i and β_i are the logit sensitivity and logit specificity for the i^{th} study, α and σ_A^2 are the mean and variance for the logit sensitivities, β and σ_B^2 are the mean and variance for the logit specificities, and σ_{AB} is the covariance between α_i and β_i across studies, respectively. In some of the literature, it is common to replace the covariance term σ_{AB} by the multiplication $\rho\sigma_A\sigma_B$ to include the correlation ρ in the model,[4] so the covariance matrix in equation (1) can be written as

$$\Sigma = \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} \tag{2}$$

Thus, the five parameters $(\alpha, \beta, \sigma_a^2, \sigma_b^2, \rho)$ need to be estimated in order to make inferences on the sensitivity and specificity.

For a test accuracy review with k studies, let TP_i, TN_i, $n_{A,i}$ and $n_{B,i}$ be the number of true positives, true negatives, diseased, and non-diseased for the i^{th} study, respectively. Chu and Cole[4] pointed out that a binomial

likelihood should be used for modelling within-study variability especially if the data are sparse, so the model should include the following components

$$TP_i \big| P_{A,i} \sim Binomial\big(n_{A,i}, P_{A,i}\big) \tag{3}$$

$$TN_i \big| P_{B,i} \sim Binomial\big(n_{B,i}, P_{B,i}\big) \tag{4}$$

where $P_{A,i}$ and $P_{B,i}$ represent the study-specific sensitivity and specificity, respectively. If both $P_{A,i}$ and $P_{B,i}$ are known, then TP_i and TN_i are assumed to follow independent binomial distributions.[4,13] In the random effects models, we assume that each study has its own test sensitivity and specificity, in other words the model includes a between-study variance component and correlation between $P_{A,i}$ and $P_{B,i}$, such that

$$g\big(P_{A,i}\big) = X_i^T\alpha + \alpha_i, \quad g\big(P_{B,i}\big) = Z_i^T\beta + \beta_i \tag{5}$$

where X_i is a vector of study-level covariates for $P_{A,i}$ and Z_i is a vector of study-level covariates for $P_{B,i}$ and both α_i, β_i are supposed to follow a bivariate normal distribution defined in equation (1). Although the logit link function $g(.)$ is commonly used in equation (5), other link functions can be applied. However, we will use the logit link function $g(.)$ and assume that $X_i = Z_i = 1$ in equation (5) throughout, so α and β will be the respective overall logit sensitivity and logit specificity.

The parameters of the bivariate generalised linear mixed effect model may be estimated by maximising the likelihood function. The log-likelihood function, $l\big(\alpha, \beta, \sigma_A^2, \sigma_B^2, \rho\big)$ for the model may be written as

$$\begin{aligned} l\big(\alpha, \beta, \sigma_A^2, \sigma_B^2, \rho\big) &= \log \prod_{i=1}^{k} p_r\big(TP_i, TN_i | n_{A,i}, n_{B,i}\big) = \sum_{i=1}^{k} \log p_r\big(TP_i, TN_i | n_{A,i}, n_{B,i}\big) \\ &= \sum_{i=1}^{k} \log \iint Bin\big(TP_i | n_{A,i};\, P_{A,i}\big) Bin\big(TN_i | n_{B,i};\, P_{B,i}\big) \phi\big(P_{A,i}, P_{B,i}; \alpha, \beta, \sigma_A^2, \sigma_B^2, \rho\big) \mathrm{d}P_{A,i} \mathrm{d}P_{B,i} \end{aligned} \tag{6}$$

where

$$Bin\big(TP_i | n_{A,i};\, P_{A,i}\big) = \binom{n_{A,i}}{TP_i} {P_{A,i}}^{TP_i} \big(1 - P_{A,i}\big)^{n_{A,i} - TP_i} \tag{7}$$

$$Bin\big(TN_i | n_{B,i};\, P_{B,i}\big) = \binom{n_{B,i}}{TN_i} {P_{B,i}}^{TN_i} \big(1 - P_{B,i}\big)^{n_{B,i} - TN_i} \tag{8}$$

and $\phi = \phi\big(P_{A,i}, P_{B,i}; \alpha, \beta, \sigma_A^2, \sigma_B^2, \rho\big)$ is the bivariate logit normal distribution, such that

$$\phi = Ke^{\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{\left(logit\left(P_{A,i}\right) - \alpha\right)^2}{{\sigma_A}^2} + \frac{\left(logit\left(P_{B,i}\right) - \beta\right)^2}{{\sigma_B}^2} - \frac{2\rho\left(logit\left(P_{A,i}\right) - \alpha\right)\left(logit\left(P_{B,i}\right) - \beta\right)}{\sigma_A \sigma_B} \right] \right\}} \tag{9}$$

where

$$K = \frac{1}{2\pi\sigma_A\sigma_B\sqrt{1-\rho^2} P_{A,i}\big(1 - P_{A,i}\big) P_{B,i}\big(1 - P_{B,i}\big)}$$

From inspecting the log likelihood function in equation (6), it can be seen that it involves a double integration over the random effects and there is no closed form so it cannot be solved analytically. In order to get a solution to the integral, we have to use numerical optimisation methods such as the Laplacian approximation or the adaptive Gaussian quadrature[14] to evaluate this integral. Before proceeding to derive the maximum likelihood estimates of the BRM using methods based on the Newton-Raphson algorithm,[7] we will briefly describe the optimisation approaches used in two generic packages.

3 Optimisation methods used in generic packages

Both the *glmer* function in the *lme4* package in R[5] and the *NLMIXED* function in SAS[6] are generic functions that have been developed to optimise a range of generalised mixed and non-linear mixed models. As such they may be

used to provide estimates for the bivariate generalised linear mixed model or BRM. Both use a Cholesky parameterisation of the models being optimised.[5,6]

Briefly, one of the issues in estimating the parameters in any generalised mixed model is that the covariance matrix of random effects, $\Sigma(\theta)$, may be singular and thus its inverse may not exist. In some cases, this may be overcome by re-formulating the objective function. Thus, for random effects vector V, $\Sigma(\theta)$ may be re-formulated in terms of a relative covariance factor $\Lambda(\theta)$, for a variance component θ, allowing V to be expressed as the product $\Lambda(\theta)U$, where U is a spherical random effects vector. Taking this approach, the likelihood function may be written in terms of sparse Cholesky factors and finding the maximum likelihood is transformed into finding the penalised least squares.[5,15] By writing the likelihood in terms of sparse Cholesky factors, the problem may be reformulated so that the resulting matrix is not singular even when $\Sigma(\theta)$ is singular.[15]

This is the approach taken in the *glmer* function in the *lme4* package in R[5] and the initial values of θ for the sparse Cholesky factors are taken to be 1 on the diagonal and 0 for off diagonal elements.[16]

The default numerical optimisation algorithms used in *glmer* are the Nelder–Mead and the Bounded Optimisation By Quadratic Approximation (BOBYQA).[17] The Nelder–Mead method is a derivative-free optimisation (DFO) algorithm[18] introduced as a means of optimising functions when the derivatives are not available or unknown. It starts with a simplex (a generalisation of a triangle to n dimensions) so that a function of n variables is evaluated at $n+1$ points. The values of the function at these points are ranked and by geometric transformations (reflection, contraction, and expansion) the point where the function is largest is replaced with a point where the function is smaller. This gives a new simplex and the process continues until convergence.

The BOBYQA algorithm is a sophisticated algorithm and one of several due to Powell which is derivative free.[19] Essentially it is based on using a quadratic model to locally approximate the objective function, F, over a trust region. After k iterations, the coefficients of the quadratic model Q_k are obtained by constraining Q_k to interpolate F at a fixed number of points – these are the interpolation conditions. The sub-problem is to find d_k such that $x_k + d_k$ minimises Q_k over the trust region. If $x_k + d_k$ improves on the current iterate x_k, then this becomes the new iterate x_{k+1} and the trust region and quadratic model Q_k are updated. If it is not an improvement, then an alternative iteration algorithm is used to identify d_k so that it ensures linear independence in the interpolation conditions. Broadly, this process continues until convergence.

Other derivative approaches may be used to fit the bivariate model as is the case with *NLMIXED* function in SAS. For instance *NLMIXED*, as used by some authors,[4,20] tends to be fitted using the default dual quasi-Newton algorithm.[6] Thus, for a symmetrical, positive definite matrix $B^{(k)}$ which satisfies the secant condition, $B^{(k)}$ is chosen so that it may be updated according to $B^{(k+1)} = B^{(k)} + A^{(k)}$ (where $A^{(k)}$ is a matrix which is easily estimated) whilst still preserving symmetry, positive definiteness and the secant condition. The Broyden, Fletcher, Goldfarb, and Shanno (BFGS) formula[21] provides one approach where these conditions are satisfied and this is applied to the Cholesky factor of the approximate Hessian as the default method in the *NLMIXED* function.

For the purpose of comparison with the Newton–Raphson algorithms that follow, we focussed on *glmer* in R which is open source and readily available.[22]

4 Maximum likelihood estimations for bivariate model using NR algorithm

Here we demonstrate two different numerical methods for deriving maximum likelihood estimates (MLE) for the parameters in the bivariate random effects model used in test accuracy meta-analysis. They are both based on the Newton–Raphson (NR) algorithm,[7] perhaps, one of the most common numerical methods used in optimisation. The NR algorithm is an iterative method for finding the roots of a differentiable function that generates a sequence of estimates which usually come increasingly close to the optimal solution. The algorithm is based on successive approximations to the solution, using Taylor's theorem to approximate the equation. It may be applied to both one-dimensional and higher dimensional problems by replacing the derivative with the gradient, and the reciprocal of the second derivative with the inverse of the Hessian matrix (see below).[23,24]

In essence, the task of maximum likelihood estimation may be reduced to a one of finding the roots to the derivatives of the log likelihood function, that is, finding $\alpha, \beta, \sigma_A^2, \sigma_B^2$ and ρ such that $\nabla l\left(\alpha, \beta, \sigma_A^2, \sigma_B^2, \rho\right) = 0$. Hence, the NR algorithm may be used to solve this equation iteratively. Suppose that $\hat{\theta}_k = (\hat{\alpha}_k, \hat{\beta}_k, \hat{\sigma}_{bk}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k)^T$ is the k^{th} estimate of the vector of true parameters $\theta = (\alpha, \beta, \sigma_a^2, \sigma_b^2, \rho)^T$ in the BRM with the log-likelihood function as given in equation (6). If we define the score statistic, $S(\hat{\theta}_k)$, as the ∇l and the Hessian matrix, $H(\hat{\theta}_k)$, such that

$$S\left(\hat{\theta}_k\right) = \left(\frac{\partial l}{\partial \hat{\alpha}_k} \quad \frac{\partial l}{\partial \hat{\beta}_k} \quad \frac{\partial l}{\partial \hat{\sigma}_{ak}^2} \quad \frac{\partial l}{\partial \hat{\sigma}_{bk}^2} \quad \frac{\partial l}{\partial \hat{\rho}_k} \right)^T \tag{10}$$

$$H\left(\hat{\theta}_k\right) = \begin{pmatrix} \frac{\partial^2 l}{\partial \hat{\alpha}_k^2} & \frac{\partial^2 l}{\partial \hat{\alpha}_k \partial \hat{\beta}_k} & \frac{\partial^2 l}{\partial \hat{\alpha}_k \partial \hat{\sigma}_{ak}^2} & \frac{\partial^2 l}{\partial \hat{\alpha}_k \partial \hat{\sigma}_{bk}^2} & \frac{\partial^2 l}{\partial \hat{\alpha}_k \partial \hat{\rho}_k^2} \\ \frac{\partial^2 l}{\partial \hat{\beta}_k \partial \hat{\alpha}_k} & \frac{\partial^2 l}{\partial \hat{\beta}_k^2} & \frac{\partial^2 l}{\partial \hat{\beta}_k \partial \hat{\sigma}_{ak}^2} & \frac{\partial^2 l}{\partial \hat{\beta}_k \partial \hat{\sigma}_{bk}^2} & \frac{\partial^2 l}{\partial \hat{\beta}_k \partial \hat{\rho}_k} \\ \frac{\partial^2 l}{\partial \hat{\sigma}_{ak}^2 \partial \hat{\alpha}_k} & \frac{\partial^2 l}{\partial \hat{\sigma}_{ak}^2 \partial \hat{\beta}_k} & \frac{\partial^2 l}{\partial \hat{\sigma}_{ak}^{2^2}} & \frac{\partial^2 l}{\partial \hat{\sigma}_{ak}^2 \partial \hat{\sigma}_{bk}^2} & \frac{\partial^2 l}{\partial \hat{\sigma}_{ak}^2 \partial \hat{\rho}_k} \\ \frac{\partial^2 l}{\partial \hat{\sigma}_{bk}^2 \partial \hat{\alpha}_k} & \frac{\partial^2 l}{\partial \hat{\sigma}_{bk}^2 \partial \hat{\beta}_k} & \frac{\partial^2 l}{\partial \hat{\sigma}_{bk}^2 \partial \hat{\sigma}_{ak}^2} & \frac{\partial^2 l}{\partial \hat{\sigma}_{bk}^{2^2}} & \frac{\partial^2 l}{\partial \hat{\sigma}_{bk}^2 \partial \hat{\rho}_k} \\ \frac{\partial^2 l}{\partial \hat{\rho}_k \partial \hat{\alpha}_k} & \frac{\partial^2 l}{\partial \hat{\rho}_k \partial \hat{\beta}_k} & \frac{\partial^2 l}{\partial \hat{\rho}_k \partial \hat{\sigma}_{ak}^2} & \frac{\partial^2 l}{\partial \hat{\rho}_k \partial \hat{\sigma}_{bk}^2} & \frac{\partial^2 l}{\partial \hat{\rho}_k^2} \end{pmatrix} \tag{11}$$

then by using Taylor's expansion of the score function $S(\hat{\theta}_k)$ we have

$$S\left(\hat{\theta}_{k+1}\right) \approx S\left(\hat{\theta}_k\right) + H\left(\hat{\theta}_k\right)\left(\hat{\theta}_{k+1} - \hat{\theta}_k\right) \tag{12}$$

Since $S\left(\hat{\theta}_{k+1}\right) = 0$ when $\hat{\theta}_{k+1}$ maximises $l_n(\theta|x_1, x_2)$, we obtain the following estimate

$$\hat{\theta}_{k+1} \approx \hat{\theta}_k - H\left(\hat{\theta}_k\right)^{-1} S\left(\hat{\theta}_k\right) \tag{13}$$

which is the k^{th} iteration of the Newton–Raphson algorithm based on the observed Fisher information (OFI) matrix (equivalent to the negative of the Hessian matrix) for estimating the five parameters in the BRM.

In order to calculate the derivatives in equations (10) and (11) numerically, one can use the simple approximation to the first order derivative in five dimensions with respect to the underlying estimated parameter. Suppose it is $\hat{\alpha}_k$, then the derivative can be approximated as

$$\frac{\partial l}{\partial \hat{\alpha}_k} = \frac{f\left(\hat{\alpha}_k + h, \hat{\beta}_k, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k\right) - f\left(\hat{\theta}_k^T\right)}{h} \tag{14}$$

or

$$\frac{\partial l}{\partial \hat{\alpha}_k} = \frac{f\left(\hat{\alpha}_k + h, \hat{\beta}_k, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k\right) - f\left(\hat{\alpha}_k - h, \hat{\beta}_k, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k\right)}{2h} \tag{15}$$

where h is very small ($h \to 0$, for example $h = 0.0001$), and $\hat{\theta}_k = (\hat{\alpha}_k, \hat{\beta}_k, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k)^T$. On the other hand, we can obtain a numerical approximation to the second-order derivative in five dimensions with respect to $\hat{\alpha}_k$ using the formula

$$\frac{\partial^2 l}{\partial \hat{\alpha}_k{}^2} = \frac{f\left(\hat{\alpha}_k + h, \hat{\beta}_k, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k\right) - 2f\left(\hat{\theta}_k^T\right) + f(\hat{\alpha}_k - h, \hat{\beta}_k, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k)}{h^2} \tag{16}$$

and the approximation to the second-order derivative in five dimensions with respect to $\hat{\alpha}_k$, $\hat{\beta}_k$ can be written as

$$\frac{\partial^2 l}{\partial \hat{\alpha}_k \partial \hat{\beta}_k} = \frac{f\left(\hat{\alpha}_k + h, \hat{\beta}_k + h, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k\right) - 2f\left(\hat{\theta}_k^T\right) + f(\hat{\alpha}_k - h, \hat{\beta}_k - h, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2, \hat{\rho}_k)}{2h^2 - \left(\frac{\partial^2 l}{\partial \hat{\alpha}_k{}^2} + \frac{\partial^2 l}{\partial \hat{\beta}_k^2}\right) \Big/ 2} \tag{17}$$

We can calculate the other elements in equations (10) and (11), in a similar fashion to those shown in equations (14) to (17). Alternatively one may use the ready-made functions in R, *grad* and *hessian*, in the package *numDeriv*.[25]

The double integration over the random effects in the log likelihood function in equation (6) is computed using the adaptive multidimensional integration algorithms described in Genz and Malik[26] and Berntsen et al.[27] It is written in C and may be accessed via the R wrapper *cubature*.[28] We can use the function *adaptIntegrate* (within *cubature*) to perform adaptive multidimensional integration of vector-valued integrands over hypercubes, and get a solution to the integral in equation (6) and then estimate the five parameters in the BRM.

The first algorithm uses the profile of the log likelihood equation[6] in equation (6) to estimate the five unknown parameters in equation (9) by starting with what may be called 'robust initial values'. The robust initial values are

starting values that are sufficiently close to the actual values of the parameters so they increase both the chances and the speed of convergence. The second algorithm is based on the observed Fisher information matrix[8] where similar to the first algorithm, robust initial values provide the starting point to the algorithm before updating the observed Fisher information matrix.

5 The method of profiling

In order to explain the method of profiling,[8,29] suppose that only two parameters α and β need to be estimated and that $\hat{\beta}$, the MLE for β, may be expressed as a function of α. The profile likelihood of α is then $L(\alpha, \hat{\beta}(\alpha))$ and is now a function of α only.[30] If $\hat{\beta}(\alpha)$ is known explicitly, then maximising the profile likelihood with respect to α is achieved easily. However, when it is not known, $\hat{\beta}(\alpha)$ may be obtained numerically by fixing α and maximising $L(\alpha, \beta)$ with respect to β. Thus $\hat{\beta}(\alpha)$ takes a different value for each fixed value of α and $\hat{\alpha}$ is the estimate for α which maximises the profile likelihood $L(\alpha, \hat{\beta}(\alpha))$. In practical terms, this means deriving profile likelihood estimates over a range of values for α and when there are more than two parameters to estimate, the range of values of the other parameters also need to be considered (see below).

Lindstrom and Bates[31] pointed out that optimising the profile log-likelihood usually requires fewer iterations, the derivatives are somewhat simpler, and the convergence is more consistent. In addition, they have also encountered examples where the NR algorithm failed to converge when optimising the likelihood (which includes a variance term) but was able to optimise the profile likelihood with ease.

It is often difficult to determine whether an algorithm has converged upon a 'local' maximum instead of the 'global' maximum[32,33] but many objective functions will have local maxima either due to the shape of the underlying function or due to noise introduced by the data. One approach to overcome this is to choose multiple initial values randomly and select the maximum these yield.[33] Here a more systematic approach is taken, where the data from the studies help define a feasible space for the global maximum and an equally spaced grid is overlaid on the space.[34,35] This is then used as the basis for a maximum likelihood approach in determining robust initial values. It represents the first phase of the algorithm. In the second phase, we update the estimations continuously, using the last estimated values, until we get the convergence.

The profile log likelihood algorithm for estimating the parameters in bivariate model:

5.1 Initial estimate phase: we can derive an initial estimate of the nuisance parameters $(\rho, \sigma_a^2, \sigma_b^2)$ by following the profile log likelihood procedure outlined above. Specifically,

5.1a. Using the minimum and maximum of α and β across all the studies as bounds, and using the delta-method to estimate the range of σ_a^2, σ_b^2, generate a regular equally-spaced sequence for each of $\sigma_a^2, \sigma_b^2, \alpha, \beta$. Next, construct a grid of all possible combinations of values of $(\sigma_a^2, \sigma_b^2, \alpha, \beta)$ where each combination of $(\sigma_a^2, \sigma_b^2, \alpha, \beta)$ generates a new log likelihood curve $l\big(\rho, \sigma_a^2(\rho), \sigma_b^2(\rho), \alpha(\rho), \beta(\rho)\big)$ over ρ. Choose the combination $(\sigma_{a,opt1}^2, \sigma_{b,opt1}^2, \alpha_{opt1}, \beta_{opt1})$ which gives the largest likelihood over all these curves when $\rho = 0$. The associated likelihood curve for this combination is then maximised with respect to ρ using the NR algorithm to give an initial estimate, $\hat{\rho}_0$.

5.1b. Construct combinations of all the possible values of $(\sigma_b^2, \alpha, \beta)$ as in 5.1a. Choose the combination of $(\sigma_{b,opt2}^2, \alpha_{opt2}, \beta_{opt2})$ which gives the largest likelihood for $\rho = \hat{\rho}_0$ and $\sigma_a^2 = \sigma_{a,opt1}^2$ from 5.1a. The associated likelihood curve for this combination with $\rho = \hat{\rho}_0$ is maximised with respect to σ_a^2 using the NR algorithm to give an initial estimate $\hat{\sigma}_{a0}^2$.

5.1c. As previously, construct combinations of all the possible values of (α, β) and choose the combination $(\alpha_{opt3}, \beta_{opt3})$ which gives the largest likelihood for $\rho = \hat{\rho}_0$, $\sigma_a^2 = \hat{\sigma}_{a0}^2$ and $\sigma_b^2 = \sigma_{b,opt2}^2$ is chosen. The associated likelihood curve for this combination with $\rho = \hat{\rho}_0$ and $\sigma_a^2 = \hat{\sigma}_{a0}^2$ is then maximised with respect to σ_b^2 using the NR algorithm to give an initial estimate $\hat{\sigma}_{b0}^2$.

5.1d. Following the same procedure, initial estimates for $\hat{\alpha}_0$ and $\hat{\beta}_0$ may be derived.

5.2. The updating phase: based on the initial estimate $\hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\sigma}_{a0}^2, \hat{\sigma}_{b0}^2, \hat{\rho}_0)^T$ from 5.1, the algorithm iteratively updates each parameter separately with the other consecutive estimated parameters. In other words, the estimate $\hat{\rho}_k$ is updated with $(\hat{\alpha}_k, \hat{\beta}_k, \hat{\sigma}_{ak}^2, \hat{\sigma}_{bk}^2)$ to get $\hat{\rho}_{k+1}$ by maximising $l(\hat{\rho}_k, \hat{\alpha}_k(\hat{\rho}_k), \hat{\beta}_k(\hat{\rho}_k), \hat{\sigma}_{ak}^2(\hat{\rho}_k), \hat{\sigma}_{bk}^2(\hat{\rho}_k))$ with respect to $\hat{\rho}_k$. Similarly, the estimate of $\hat{\sigma}_{ak}^2$ is updated with $(\hat{\alpha}_k, \hat{\beta}_k, \hat{\sigma}_{bk}^2, \hat{\rho}_{k+1})$ to get $\hat{\sigma}_{ak+1}^2$, $\hat{\sigma}_{bk}^2$ with $(\hat{\alpha}_k, \hat{\beta}_k, \hat{\sigma}_{ak+1}^2, \hat{\rho}_{k+1})$ to get $\hat{\sigma}_{bk+1}^2$, $\hat{\alpha}_k$ with $(\hat{\beta}_k, \hat{\sigma}_{ak+1}^2, \hat{\sigma}_{bk+1}^2, \hat{\rho}_{k+1})$ to get $\hat{\alpha}_{k+1}$, and $\hat{\beta}_k$ with $(\hat{\alpha}_{k+1}, \hat{\sigma}_{ak+1}^2, \hat{\sigma}_{bk+1}^2, \hat{\rho}_{k+1})$ to get $\hat{\beta}_{k+1}$. So, at the end of this process we have $\hat{\theta}_{k+1} = (\hat{\alpha}_{k+1}, \hat{\beta}_{k+1}, \hat{\sigma}_{ak+1}^2, \hat{\sigma}_{bk+1}^2, \hat{\rho}_{k+1})$.

5.3. While $|\hat{\theta}_{k+1} - \hat{\theta}_k| > \varepsilon$, set $k = k + 1$ and repeat 5.2 until convergence is achieved.

Although the algorithm is straightforward, compared with the observed Fisher information algorithm below, it is more computationally expensive and is likely to be more time consuming as a result. In particular, the second phase involves several iterations, as the NR algorithm is applied to each of the five parameters individually in each update until convergence is achieved. Moreover, the log likelihood function is evaluated over many different possible combinations of the parameters' values.

6 Observed Fisher information with robust initial values (OFIRIV)

Although the method of profiling circumvents the local maximum problem by generating robust initial parameter values, it is computationally expensive. In contrast, the observed Fisher information is more efficient than the method of profiling but without appropriate starting values there is still the risk of it converging on a local maximum.

Here the approach of ascertaining robust initial parameter values is combined with an algorithm based on the observed Fisher information.[8] This has the potential of improving on the previous algorithm by increasing the computational efficiency.

Thus, the algorithm is as follows:

6.1. Initial estimate phase: get an initial estimate $\hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\sigma}^2_{a0}, \hat{\sigma}^2_{b0}, \hat{\rho}_0)^T$ for the parameters $(\alpha, \beta, \sigma^2_a, \sigma^2_b, \rho)$ by using the algorithm described in 5.1a to 5.1d

6.2. Updating phase: the next steps use the observed Fisher information matrix[8] to update the estimates for the parameters in the BRM.

6.2a. Let $\theta = (\alpha, \beta, \sigma^2_a, \sigma^2_b, \rho)^T$ be the vector of parameters to be estimated in the BRM with log-likelihood function defined in equation (6), set $k = 0$ and choose the initial value $\hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\sigma}^2_{a0}, \hat{\sigma}^2_{b0}, \hat{\rho}_0)^T$ from 6.1 to start the algorithm.

6.2b. Calculate the score statistic $S(\hat{\theta}_k)$ and the Hessian matrix $H(\hat{\theta}_k)$ as in equations (10) and (11), respectively.

6.2c. Estimate $\hat{\theta}_{k+1}$ based on $\hat{\theta}_k$ such that: $\hat{\theta}_{k+1} = \hat{\theta}_k - [H(\hat{\theta}_k)]^{-1}\, S(\hat{\theta}_k)$.

6.2d. Check whether $\hat{\theta}_{k+1}$ is optimal using the convergence condition $|\hat{\theta}_{k+1} - \hat{\theta}_k| \leq \varepsilon$, where ε expresses the desired tolerance level and is usually very small, for example $\varepsilon = 10^{-12}$.

6.2e. While $|\hat{\theta}_{k+1} - \hat{\theta}_k| > \varepsilon$, set $k = k + 1$ and repeat 6.2b to 6.2d until we get convergence.

To ensure stability of the algorithm, we may control for jumps in individual components of the parameter vector between iterations and redirect the algorithm to the robust initial value for the component. For example, if the difference $\left|\hat{\alpha}_{k+1} - \hat{\alpha}_k\right|$ between successive iterations is too large, then we may reset $\hat{\alpha}_{k+1}$ to $\hat{\alpha}_0$.

Other criteria may be used for terminating the iteration. Recall that obtaining the maximum likelihood estimate is equivalent to finding the roots to the score statistic $S(\hat{\theta}_k)$, then a suitable stopping criterion would be when $|S(\hat{\theta}_k)^2 \leq \varepsilon|$. Alternatively we may use $-[H(\hat{\theta}_k)]^{-1}\, S(\hat{\theta}_k)^2 \leq \varepsilon$ – asymptotically the observed Fisher information is equivalent to the variance of the score statistic, so this criterion has the advantage of being insensitive to scaling of the variables. Occasionally, a parameter estimate may recur, so that $\hat{\theta}_k$ is exactly equal to $\hat{\theta}_{k+m}$ for $m > 1$. At this point, the algorithm has entered a limit cycle and a stopping rule is required so that it does not continue indefinitely.

Compared to the profile log likelihood algorithm, this algorithm consumes less time than the former and is computationally more straightforward. Furthermore, once the Hessian matrix has been estimated at the initial step, $(H(\hat{\theta}_0))$, this may be used for subsequent iterations thereby saving computation time. However, if the Hessian matrix is estimated at each iteration then the algorithm will converge after fewer iterations than if $H(\hat{\theta}_0)$ is used throughout, but will nonetheless take longer on average and the risk of getting a singular matrix will be much higher which leads to a lower convergence rate. Here $H(\hat{\theta}_0)$ was used as the estimate of the Hessian for each iteration.

It is well known that the choice of initial values can be important in the speed of convergence, the ability of the algorithm to find a global maximum, and the ability to converge at all.[36, 37] However, specifically for Newton–Raphson-based methods, Kantorovich's theorem provides the theoretical underpinning for the importance of the choice of initial values and the success of convergence.[38] Essentially around the start point, the behaviour of the Jacobian of the function and its inverse have to meet certain conditions on continuity and boundedness if the algorithm is to converge.

Here we applied a grid across a bounded space for the parameters[29] before taking a maximum likelihood approach to generate robust initial values for the parameters. However, there is no guarantee the algorithm with robust initial values will produce parameter estimates that uniquely maximises the log-likelihood. Whilst the choice of robust initial values may lower the risk of the algorithm converging on a local maximum,[39] it cannot

eliminate this risk. Essentially identifying the global maximum is still a heuristic process no matter what initial values are chosen.

Furthermore when the data are noisy, rather than converging on a local maximum, the algorithms may fail to converge at all. Generally, this occurs when one or more elements in the score function or Hessian returns an infinity, the absolute value of the correlation exceeds 1, or a negative variance begins to emerge. To cope with these types of situation, we may reset the variable responsible to either the value in a previous iteration or to the initial value. If this occurs in the initial value estimate phase, the resetting of the variable may involve setting the value on the grid that maximises the likelihood. If the correlation is the problem variable in the initial estimate phase, the Pearson correlation coefficient for the observed data may be used. These measures allow the algorithm to proceed on a slightly modified trajectory. Both algorithms discussed in this and the preceding section accommodate these scenarios in this way.

An alternative approach for obtaining the MLEs of the parameters is to transform all or part of the model in order to facilitate convergence. This is used by the two generic packages as discussed in section 3.

7 Numerical examples

In this section, the two algorithms are evaluated through a simulated study before applying them to two real case examples. In each case, they are compared with the *glmer* function from the package *lme4* in R,[5] which has been previously validated. All analyses were conducted in R[22] and the code for each of the algorithms appears in the online appendices.

7.1 Simulation study

For the simulation study, the true values of the five parameters were set to: $\alpha = 1.2$; $\beta = 2.5$; $\sigma_a^2 = 0.4$; $\sigma_b^2 = 0.6$; and $\rho = -0.7$. The number of studies k included in the meta-analysis was set at 10 and 20. Thus, the logit sensitivity α_i and logit specificity β_i for the i^{th} study were simulated from

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 1.2 \\ 2.5 \end{pmatrix}, \quad \begin{pmatrix} 0.4 & -0.3429 \\ -0.3429 & 0.6 \end{pmatrix} \right) \tag{18}$$

This provides the study-specific sensitivity, $P_{A,i} = logit^{-1}(\alpha_i)$ and specificity $P_{B,i} = logit^{-1}(\beta_i)$. For each study i, the number of non-diseased $n_{B,i}$ was generated randomly to be between 10 and 200 and the diseased $n_{A,i}$, chosen to be $0.25n_{B,i}$ rounded to the nearest whole number. Thus, for each of k studies, the true positives TP_i, and true negatives TN_i were simulated from the binomial distributions detailed in equations (3) and (4).

For each of the three algorithms (including *glmer*) the BRM was applied to 10,000 simulated data sets of size $k = 10$ and then $k = 20$. The results were compared using the convergence rate, mean squared error (MSE), average relative error (ARE), mean bias and coverage probability.

Table 1 gives the convergence rates (CR) for each of the three algorithms. It is clear that the *glmer* function does not converge for all the datasets achieving at most 84% for $k = 20$ studies. This contrasts the profile likelihood and OFIRIV methods which both have near 100% convergence. Also increasing the number of studies improves convergence for all the methods.

As heterogeneity is one of the factors contributing to non-convergence, restricting the analysis to the converged data sets potentially may make the overall sample less heterogeneous. Thus, we may expect the mean square errors

Table 1. The convergence rates calculated from 10,000 simulations for each method at $k = 10$ and 20.

	Convergence Rate	
Method	k = 10	k = 20
glmer	0.8026	0.8365
Profile likelihood	1.0000	1.0000
OFIRIV	0.9888	0.9951

(MSE) and average relative error (ARE) to be lower for the *glmer* function, where the converged set was 15% smaller than the other two algorithms. This is observed in Tables 2 and 3 below, although the differences are small.

The mean bias of the estimated values of the five parameters for each of the four methods is given in Table 4. Similar to the previous tables, the results are comparable across the different methods with no one method giving a consistently better performance over all five parameters.

Table 5 shows the coverage probabilities of the confidence ellipses for (α, β) as estimated using methods previously described.[40] The method of profiling produces the highest coverage probability for both cases.

It is clear that the different methods are comparable across a number of statistics. However, the *glmer* function does have a substantially lower convergence rates than the other two algorithms. Thus based on its superior convergence rate and coverage probability, the profile likelihood is recommended as the method of choice for estimating the parameters for the bivariate random effects model in meta-analysis.

To illustrate the contrasting performance, three examples where *glmer* failed to converge are compared with the profile likelihood and the OFIRIV algorithms which did converge. The three simulated data sets are based on 10 studies and may be found in the online Appendix. For the first example, *glmer*'s failure to converge was due to it calculating an inconsistent gradient value in some iterations ($\max|\text{grad}| = 0.0105486$). For this example, the profile likelihood estimates of ($\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$, $\hat{\rho}$) converged after five iterations to (1.2696185, 2.3511021, 0.4482178, 0.4121124, −0.4127191) and the OFIRIV converged after nine iterations to (1.3075654, 2.3844844, 0.5144346, 0.4182635, −0.4127191).

Table 2. MSE of the estimated values of the five parameters for the different methods at $k = 10$ and 20 based on converged samples from 10,000 simulations. The values in bold refer to the lowest MSE between all the methods.

	$k = 10$					$k = 20$				
Method	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$
glmer	**0.0727**	0.0859	**0.1020**	**0.1395**	0.1834	**0.0343**	**0.0409**	**0.0479**	**0.0668**	0.0660
Profile	0.0778	**0.0840**	0.1090	0.1516	0.1190	0.0359	0.0432	0.1013	0.0944	**0.0465**
OFIRIV	0.0808	0.1133	0.2620	0.1530	**0.1087**	0.0421	0.0596	0.1059	0.1537	0.0542

Table 3. ARE of the estimated values of the five parameters for the different methods at $k = 10$ and 20 based on converged samples from 10,000 simulations.

	$k = 10$					$k = 20$				
Method	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$
glmer	**0.1784**	0.0929	0.5956	**0.4847**	0.4427	**0.1227**	**0.0640**	**0.4350**	**0.3447**	0.2879
Profile	0.1792	**0.0917**	**0.5924**	0.4854	**0.3163**	0.1253	0.0656	0.5041	0.3529	**0.2212**
OFIRIV	0.1848	0.0963	0.6475	0.4975	0.3167	0.1342	0.0698	0.5675	0.3816	0.2347

Note: The values in bold refer to the ARE between all the methods.

Table 4. Mean bias of the estimated values of the five parameters for the different methods at $k = 10$ and 20 based on converged samples from 10,000 simulations.

	$k = 10$					$k = 20$				
Method	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$
glmer	**0.0106**	−0.0114	−0.0127	**−0.0518**	**0.0001**	**0.0027**	−0.0089	−0.0339	−0.0445	**−0.0257**
Profile	0.0107	0.0051	**−0.0107**	−0.0604	0.0842	0.0098	0.0041	**0.0311**	**−0.0235**	0.0330
OFIRIV	0.0108	**0.0045**	0.0158	−0.0573	0.1006	0.0094	**0.0018**	0.0658	−0.0032	0.0524

Note: The values in bold refer to the lowest absolute bias between all the methods.

Table 5. The coverage probability of the 95% confidence regions for (α, β) based on the converged samples from 10,000 simulations for each method at $k = 10$ and 20.

	Coverage probability	
Method	$k = 10$	$k = 20$
glmer	0.9442	0.9359
Profile likelihood	0.9483	0.9396
OFIRIV	0.9457	0.9395

Table 6. The estimation results (in logit space) based on the different algorithms for the CT dataset.

Algorithm	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$	Iterations
Profile log likelihood	0.6254442	1.8819420	0.2793882	0.1736081	−0.7742788	10
OFIRIV	0.6254440	1.8819419	0.2793883	0.1736061	−0.7742770	15
glmer	0.6256095	1.8821715	0.2766782	0.1728707	−0.778121	205

Note: For glmer this is the number of iterations of the Nelder–Mead algorithm.

In the second example, *glmer* returned a NAN for the correlation coefficient and two warning messages. The first was that it was unable to evaluate a scaled gradient and the second that there was a degenerate Hessian matrix with negative eigenvalues. The profile likelihood estimates of $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}_a^2, \hat{\sigma}_b^2, \hat{\rho})$ converged after four iterations to (1.6202820, 2.3936771, 0.1321239, 0.5772051, −0.1337343) and the OFIRIV algorithm converged after six iterations to (1.6266015, 2.3164244, 0.1321239, 0.5618353, −0.1624175).

In the third example, *glmer* failed to converge due to producing a correlation coefficient $\hat{\rho} = -1$ which makes equation (9) undefined. Furthermore, the algorithm gave an inconsistent gradient value for some iterations ($\max|\text{grad}| = 0.00106003$). In contrast, the profile likelihood estimates of $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}_a^2, \hat{\sigma}_b^2, \hat{\rho})$ converged after nine iterations to (1.0629164, 2.5540033, 0.3010341, 0.6466069, −0.7733131) and the OFIRIV algorithm converged after 962 iterations to (1.18282870, 2.59699272, 0.03534702, 0.19803168, −0.77331306).

7.2 Real data examples

In this section, the three algorithms described are applied to two previously published test accuracy reviews.[41,42] For each of these reviews, the five parameters in the BRM in equation (6) were estimated by the three algorithms and their performances compared.

7.2.1 Computed tomography of the distant metastasis

The first review evaluated the accuracy of several imaging modalities in detecting cancer including 98 studies published between 1990 and 2009.[41] Here the focus will be on the accuracy of computed tomography (CT) in identifying distant metastases where there were 12 relevant studies. The data may be found in the supplementary materials of Chen et al.[13]

In Table 6, the estimates of the five parameters in logit space for each of the algorithms are given for the CT data. The number of iterations required to achieve convergence by each algorithm is also given. In general, the estimated values produced from profile likelihood and the OFIRIV algorithms are very close to those estimated by the *glmer* function.

As point of illustration, Tables 7 and 8 give the successive estimates for α, β, σ_a^2, σ_b^2 and ρ for the profile log likelihood and OFIRIV algorithms at each iteration. As may be seen from both tables, the robust initial values for the profile likelihood and the OFIRIV are within a close proximity of the final estimates for the parameters. This enables more rapid convergence and reduces the risk of converging on a local maximum. Convergence is achieved after 10 iterations for the profile likelihood algorithm and 15 iterations for the OFIRIV algorithm. In general, *glmer* requires a greater number of iterations before the convergence conditions are satisfied.

Also of note is the behaviour of each algorithm which shows smooth changes between iterations without any wild fluctuations. This is because the algorithms start with robust initial values that are sufficiently close to the real value of the parameters thereby increasing the stability of the algorithms.

Table 7. Estimates for α, β, σ_a^2, σ_b^2 and ρ (in logit space) at each iteration for the profile log likelihood algorithm for the CT dataset.

Iteration	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$
RIV	0.5799135	1.8898890	0.2555894	0.2135463	−0.5987675
1	0.6158106	1.8817761	0.2708390	0.1809543	−0.7216847
2	0.6237156	1.8815193	0.2767653	0.1752273	−0.7605525
3	0.6237154	1.8822370	0.2788309	0.1739427	−0.7708735
4	0.6251286	1.8819709	0.2789662	0.1737936	−0.7731979
5	0.6253686	1.8819368	0.2792944	0.1736791	−0.7738299
6	0.6253686	1.8819368	0.2793657	0.1736244	−0.7741466
7	0.6254275	1.8819371	0.2793582	0.1736228	−0.7741466
8	0.6254413	1.8819399	0.2793822	0.1736115	−0.7742438
9	0.6254440	1.8819415	0.2793869	0.1736086	−0.7742722
10	0.6254442	1.8819420	0.2793882	0.1736081	−0.7742788

Note: RIV are the robust initial values that enter the updating part of the algorithm.

Table 8. Estimates for α, β, σ_a^2, σ_b^2 and ρ (in logit space) at each iteration for the OFIRIV algorithm for CT data. RIV are the robust initial values that enter the updating part of the algorithm.

iteration	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$
RIV	0.5799135	1.8898890	0.2555894	0.2135463	−0.5987675
1	0.6235215	1.8828534	0.2694258	0.1723672	−0.7705554
2	0.6243623	1.8819306	0.2779604	0.1743782	−0.7730370
3	0.6252492	1.8819956	0.2790256	0.1733277	−0.7736512
4	0.6254016	1.8819652	0.2793110	0.1738137	−0.7744848
5	0.6254305	1.8819280	0.2793691	0.1734812	−0.7740802
6	0.6254439	1.8819544	0.2793845	0.1736935	−0.7743981
7	0.6254419	1.8819342	0.2793871	0.1735523	−0.7741988
8	0.6254449	1.8819475	0.2793882	0.1736450	−0.7743325
9	0.6254433	1.8819386	0.2793882	0.1735839	−0.7742451
10	0.6254445	1.8819445	0.2793884	0.1736241	−0.7743028
11	0.6254437	1.8819406	0.2793883	0.1735976	−0.7742648
12	0.6254442	1.8819432	0.2793883	0.1736150	−0.7742898
13	0.6254439	1.8819415	0.2793883	0.1736035	−0.7742733
14	0.6254441	1.8819426	0.2793883	0.1736111	−0.7742842
15	0.6254440	1.8819419	0.2793883	0.1736061	−0.7742770

7.2.2 Screening for depression based on the PHQ-9

The second dataset used is a review which evaluated the accuracy of the Patient Health Questionnaire (PHQ-9) in screening for depression. The PHQ-9 consists of nine questions and is a recognised screening tool for depression. Willis and Hyde[42] conducted a meta-analysis which evaluated its accuracy and the data used here may be found in the supplemental appendix.[42] There were 10 included studies.

For each algorithm, Table 6 gives the estimated values of the five parameters for the PHQ-9 data and the number of iterations needed for convergence. Like the previous example, the OFIRIV algorithm and profile log likelihood algorithm give results that are close to those from the *glmer* function. Although the OFIRIV executes more iterations than the profile likelihood before convergence is attained, it still executes far fewer than the *glmer* function.

8 Discussion

Meta-analysis is integral to evidence synthesis providing a means of summarising research from multiple primary studies. Its widespread uptake has coincided with developments in the meta-analysis methods used, progressing from fixed effects methods[43] to including study-specific random effects,[44] and from univariate outcomes[44] to using multivariate outcomes.[45]

Table 9. The estimation results (in logit space) based on the different algorithms for the PHQ-9 dataset.[42]

Algorithm	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\rho}$	Iterations
Profile log likelihood	1.0575056	2.3793688	0.4784003	0.6340357	−0.5801280	7
OFIRIV	1.0575050	2.3793687	0.4784016	0.6340387	−0.5801333	57
glmer	1.057142	2.379097	0.4742536	0.6313224	−0.5837202	212

Note: For *glmer* this is the number of iterations of the Nelder–Mead algorithm.

This has increased the complexity of the type of models used and the optimisation methods needed to estimate the unknown parameters. The most common model used in test accuracy meta-analyses is a bivariate generalised linear mixed model, and is often referred to as the bivariate random effects model (BRM). The complexity of this model lies with the need to perform a double integration over the random effects and an integrand which is a binomial-normal mixture distribution. Having no closed form, numerical methods are required to estimate the parameters of interest. Although generic functions such as *glmer* in the *lme4* package in R[5] and *NLMIXED* in SAS[6] may be used to fit the BRM, they remain 'black boxes' to the vast majority of users.

Here we have demonstrated from first principles how maximum likelihood estimates may be derived using Newton–Raphson-based approaches to provide estimates for the parameters of interest in the BRM used in test accuracy meta-analyses. In this respect, the proposed algorithms appear to have received little attention in the literature.

Both the method of profiling and the Observed Fisher Information matrix algorithm perform well and give accurate estimates for the five unknown parameters of the BRM. However, without suitable modifications, they still have the potential to breakdown either by converging on biased estimates, the so-called 'local maxima problem',[39] or not converge at all.

One way to address the local maxima problem is to choose the initial values for the parameters more carefully. Here we get robust initial values by first using the data to derive a grid across a feasible space of values for the parameters. Then each parameter is estimated independently based on values of the other parameters that maximise the log likelihood function with respect to the parameter being estimated. This method is aimed at providing initial values which are close to the true values for the parameters to increase the chances of converging on these true values.

The second issue is that the algorithm may fail to converge at all, particularly when there are noisy data. There may be a number of reasons for this, including difficulty in calculating the partial second derivatives in the Hessian matrix due to their being a very small rate of change or that an inverse for the Hessian matrix may not exist. The correlation may become out of bounds or one or more of the variances may take on negative values. Essentially this represents a recurring challenge for multi-parameter models – how to ensure the optimisation algorithm reliably converges on an accurate estimate.

To deal with this, some authors advocate transforming the model to an alternative parameterisation such as those used by the generic packages discussed earlier. For example, the model may be transformed so that the covariance matrix or Hessian matrix remains positive definite throughout successive iterations. Whilst this offers a substantial improvement, for the *glmer* function at least, it does not lead to convergence in all cases. This was clearly demonstrated by the simulation study.

Another approach is to monitor the iterative process for aberrant parameter estimates or function values and reset to a value from a previous iteration when this occurs. For example, when a parameter estimate strays out of the space of feasible values, or a derivative becomes infinite. This recognises there may be many trajectories that converge on a stable estimate and resetting the current estimate of a parameter may move the algorithm onto a different trajectory. This was the method used in both the profile likelihood and the OFIRIV algorithms and the convergence rates were 100% and close to 100%, respectively.

Both algorithms developed in this study perform better than the *glmer* function in terms of convergence and coverage probability whilst being comparable in other performance characteristics such as mean squared error, mean bias and average relative error. However, due to its superior convergence rate and coverage probability, we recommend the method of profiling over the OFIRIV.

Furthermore the OFIRIV and method of profile algorithms benefit from having been developed specifically to estimate the parameters in the BRM, in contrast to the *glmer* function which is designed to fit a range of different models. Perhaps this indicates that as the models get more sophisticated, algorithms which are specifically optimised for the task may become more important.

Other Newton–Raphson-based approaches are possible, such as the method of scoring which uses the expected Fisher information matrix.[46] In principle, this method should improve the stability of the algorithm by ensuring the Hessian matrix is positive definite. However, for the BRM it involves two integrations, one over the random effects and the other to estimate the expectation of the Hessian matrix and technically this is not straight forward as well as being computationally time-consuming.

Although the focus here has been on developing algorithms which estimated the sensitivity and specificity in a BRM, the same approach could easily be extended to estimating parameters when study-level covariates are included in the BRM. Such meta-regression analyses are common place when investigating heterogeneity between studies and may improve the potential validity of any estimates.[47] Equally the algorithms could be applied to recently developed tailored models which augment the applicability of test accuracy research by combining meta-analyses with routine data.[48,49]

The study does have some limitations. Although the OFIRIV and method of profiling algorithms demonstrate high performance characteristics and compare favourably with the one of the generic functions in R, a more extensive investigation is required to firmly establish their utility and limitations. This would involve evaluating them over a greater variety of cases, including examples with sparse data.[50]

Many of the functions used to fit the BRM invoke generic optimisation methods[5,6] that are used to fit other models. For example, *glmer* uses Nelder–Mead[18] and BOBYQA[19] and *NLMIXED* uses a dual quasi-Newton algorithm[6] as the default algorithm across all types of models. One of the conclusions which may be drawn from this study is that it may be for the BRM a more specific optimisation approach would overcome some of the convergence issues that have been previously reported in other studies.[50] This could be investigated using simulated examples over a range of optimisation algorithms.

The emphasis here has been to be explicit in the methods used to fit the bivariate random effects model and demonstrate how this may be done from first principles using the open source programming language R.[22] However, as an interpretative language, R is slow for such models and the code may take several minutes to run. The computational time could be significantly improved by translating the algorithms into a low-level compiled language such as C.

In summary, we have developed two algorithms based on Newton–Raphson methods to fit specifically the bivariate random effects model used in meta-analysis of test accuracy studies. From a simulation study, it was demonstrated that both algorithms had higher convergence rates and coverage probability than those from the *glmer* function whilst having similar performance characteristics in other measures. Overall the profile likelihood approach had the best performance characteristics for fitting the bivariate random effects model out of the three methods. Future research should focus on improving the computational time of these algorithms.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: BHW was supported by funding from a Medical Research Council Clinician Scientist award (MR/N007999/1).

Supplemental material

Supplemental material for this article is available online.

ORCID iD


Brian H Willis https://orcid.org/0000-0002-0821-8624

References

1. Stampfer MJ, Goldhaber SZ, Yusuf S, et al. Effect of intravenous streptokinase on acute myocardial infarction: pooled results from randomized trials. *N Engl J Med* 1982; **307**: 1180–1182.

2. Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992; **327**: 248–254.
3. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; **58**: 982–990.
4. Chu H and Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol* 2006; **59**: 1331.
5. Bates D, Maechler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015; **67**: 1–48.
6. SAS Institute Inc. *SAS/STAT® 14.1 User's guide*: Chapter 82. Cary, NC: SAS Institute Inc, 2015, https://support.sas.com/documentation/onlinedoc/stat/141/nlmixed.pdf (accessed 15 January 2019).
7. Givens GH and Hoeting JA. *Computational statistics*. 2nd ed. New Jersey: John Wiley and Sons, 2013, pp.26-38.
8. Givens GH and Hoeting JA. *Computational statistics*. 2nd ed. New Jersey: John Wiley and Sons, 2013: 9-11.
9. Krzanowski WJ and Hand DJ. *ROC curves for continuous data*. London: CRC Press, 2009, pp.19-30.
10. Moses LE, Shapiro D and Littenberg B. Combining independent studies of a diagnostic test into a summary roc curve: data-analytic approaches and some additional considerations. *Stat Med* 1993; **12**: 1293–1316.
11. Rutter CM and Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2005; **20**: 2865–2884.
12. Van Houwelingen HC, Zwinderman KH and Stijnen T. A bivariate approach to meta-analysis. *Stat Med* 1993; **12**: 2273–2284.
13. Chen Y, Liu Y, Ning J, et al. Composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Stat Meth Med Res* 2017; **26**: 914–930.
14. Pinheiro JC and Chao EC. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J Comput Graph Stat* 2006; **15**: 58–81.
15. https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf (accessed 23 May 2019).
16. https://stackoverflow.com/questions/39394110/mixed-model-starting-values-for-lme4 (accessed 23 May 2019).
17. https://www.rdocumentation.org/packages/lme4/versions/1.1-18-1/topics/lmerControl (accessed 23 May 2019).
18. Nelder JA and Mead R. A simplex algorithm for function minimization. *Computer J* 1965; **7**: 308–313.
19. Powell MJD. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report, Department of Applied Mathematics and Theoretical Physics, Cambridge University, Cambridge, England. 2009.
20. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidem* 2004; **57**: 925–932.
21. Byrd RH, Lu P, Nocedal J, et al. A limited memory algorithm for bound constrained optimization. *SIAM J Scientific Comput* 1995; **16**: 1190–1208.
22. R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018, http://www.R-project.org
23. Griva I, Nash SG and Sofer A. *Linear and nonlinear optimization*. Virginia: SIAM, 2009, pp.67-74.
24. Nocedal J and Wright SJ. *Numerical optimization*. 2nd ed. New York, NY: Springer, 2006, pp.22-29.
25. Gilbert P and Varadhan R. numDeriv: Accurate Numerical Derivatives. R package version 2016.8-1, https://CRAN.R-project.org/package = numDeriv
26. Genz AC and Malik AA. An adaptive algorithm for numeric integration over an N-dimensional rectangular region. *J Comput Appl Math* 1980; **6**: 295–302.
27. Berntsen J, Espelid TO and Genz A. An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Trans Math Soft* 1991; **17**: 437–451.
28. Narasimhan B and Johnson SG. Cubature: Adaptive multivariate integration over hypercubes. 2017 R package version 1.3-11, https://CRAN.R-project.org/package=cubature
29. Cole SR, Chu H and Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am J Epidemiol* 2014; **179**: 252–260.
30. Murphy SA and van der Vaart AW. On profile likelihood. *J Am Stat Assoc* 2000; **95**: 449–465.
31. Lindstrom MJ and Bates DM. Newton-Raphson and EM algorithms for linear mixed effects models for repeated-measures data. *J Am Stat Assoc* 1988; **83**: 1014–1022.
32. Gentle JE. *Computational statistics*. London: Springer, 2009, p.243.
33. Givens GH and Hoeting JA. Computational statistics. 2nd ed. New Jersey: John Wiley and Sons, 2013, p.66.
34. Laird N. Nonparametric maximum likelihood estimation of a mixing distribution. *J Am Stat Assoc* 1978; **73**: 805–811.
35. Edgar TF, Himmelblau DM and Lasdon L. *Optimization of chemical processes*. 2nd ed. New York: McGraw-Hill, 2001, p.183.
36. Karlis D and Xekalaki E. Choosing initial values for the EM algorithm for finite mixtures. *Computat Stat Data Analys* 2003; **41**: 577–590.
37. Gertz M, Nocedal J and Sartenaer A. A starting-point strategy for non-linear interior methods. *Appl Math Lett* 2004; **17**: 945–952.
38. Tapia R. The kantorovich theorem for Newton's method. *Am Math Monthly* 1971; **78**: 389–392.
39. Myung IJ. Tutorial on maximum likelihood estimation. *J Math Psychol* 2003; **47**: 90–100.

40. Harbord RM, Deeks JJ, Egger M, et al. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; **8**: 239–51.
41. Xing Y, Bronstein Y, Ross MI, et al. Contemporary diagnostic imaging modalities for the staging and surveillance of melanoma patients: a meta-analysis. *J Natl Cancer Inst* 2011; **103**: 129–142.
42. Willis BH and Hyde CJ. What is the test's accuracy in my practice population? Tailored meta-analysis provides a plausible estimate. *J Clin Epidemiol* 2015; **68**: 847–854.
43. Mantel N and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22**: 719–748.
44. DerSimonian R and Laird N. Meta-analysis in clinical trials. *Control Clin trials* 1986; **7**: 177–188.
45. Van Houwelingen HC, Arends LR and Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002; **21**: 589–624.
46. Wand M. Fisher information for generalised linear mixed models. *J Multivar Anal* 2007; **98**: 1412–1416.
47. Willis BH and Riley RD. Measuring the statistical validity of summary meta-analysis and meta-regression results for use in clinical practice. *Stat Med* 2017; **36**: 3283–3301.
48. Willis BH and Hyde CJ. Estimating a test's accuracy using tailored meta-analysis – How setting-specific data may aid study selection. *J Clin Epidemiol* 2014; **67**: 538–546.
49. Willis BH, Coomar D and Baragilly M. Tailored meta-analysis: an investigation of the correlation between the test positive rate and prevalence. *J Clin Epidemiol* 2019; **106**: 1–9.
50. Takwoingi Y, Guo B, Riley RD, et al. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Meth Med Res* 2017; **26**: 1896–1911.